Mining Serial Transcriptomes for Modularly Expressed Genes to Aid Dynamic Understanding of Gene Functions.

ZhiLiang Ji (纪志梁)
Web: http://bioinf.xmu.edu.cn
Email: appo@xmu.edu.cn

厦门大学生物信息学辅助药物开发研究组

Bioinformatics-Aided Drug Discovery Group, Xiamen University

Part I: Data Mining & Systems Biology

数据挖掘与系统生物学

通过对高通量技术如Microarray、新一代测序技术（NGS）等的数据挖掘，在全基因组的大尺度，开展蛋白质或基因家族的功能分析，及系统生物学研究。

PaGeFinder Pattern Gene Finder

bioinf.xmu.edu.cn

HOME | PaGenBase | MFE | DITOP |

HOME | ANALYSIS | DOWNLOAD | HELP

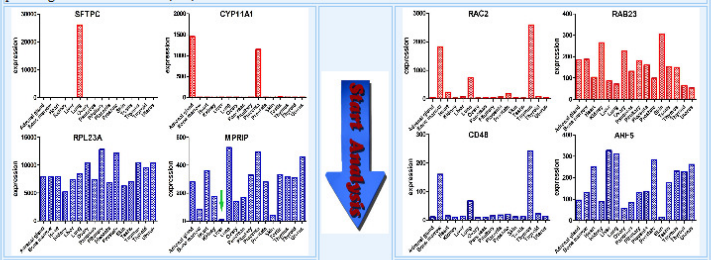
Pattern Gene Finder

Last Update 2012.02

About PaGeFinder

Pattern Gene Finder (PaGeFinder) is a free on-line server to provide interactive pattern analysis of user-submitted gene expression profiles generated by high throughput technologies, e.g. microarray, RNA-seq, SAGE. PaGeFinder implements new algorithms and functions for quantitative identification of pattern genes like specific/selective genes, housekeeping genes and repressed genes. PaGeFinder is particularly useful for dynamic analysis of serial transcriptomic data under different spatiotemporal conditions (tissues, developmental stages, physiological states, etc). It would be also useful in comparing gene expression profiles from the same platform or even different platforms.

To cite: Pan, J.B., Hu, S.C., Wang, H., Zou, Q. and Ji, Z.L. (2012) PaGeFinder: quantitative identification of spatiotemporal pattern genes. Bioinformatics, 28, 1544-1545.

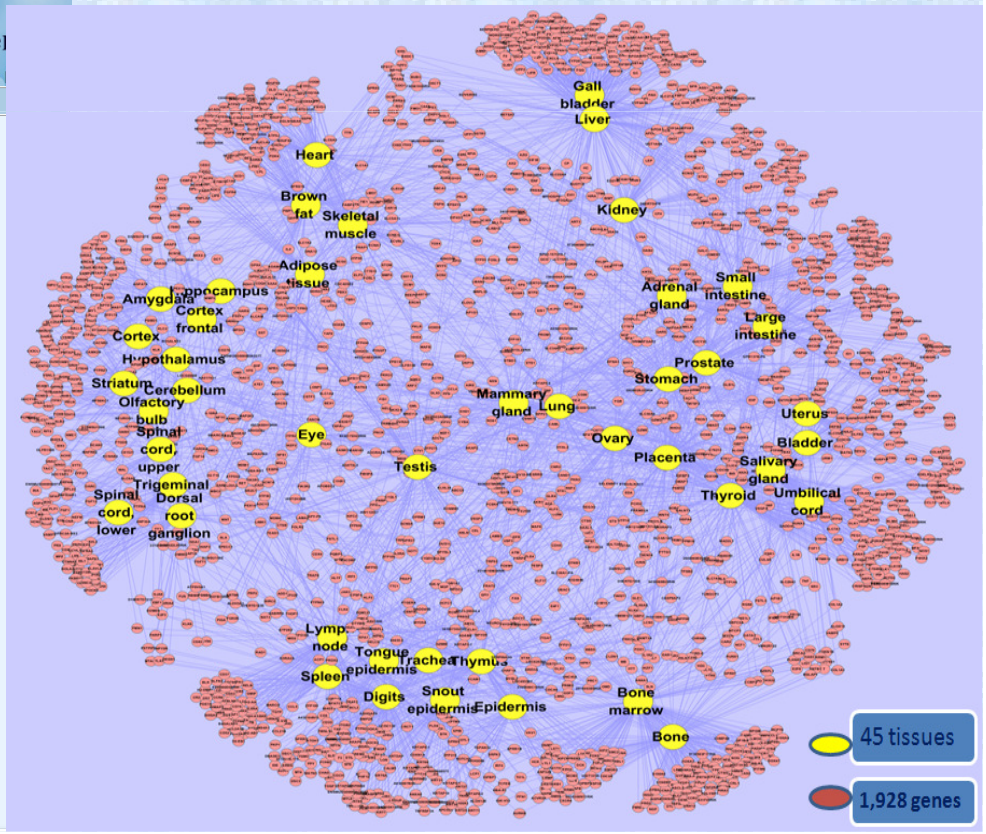


Pattern Gene Search | *Similarity / Correlation Analysis*

Useful Links

TiSGeD - a database for quantitative identification of tissue-specific gene.

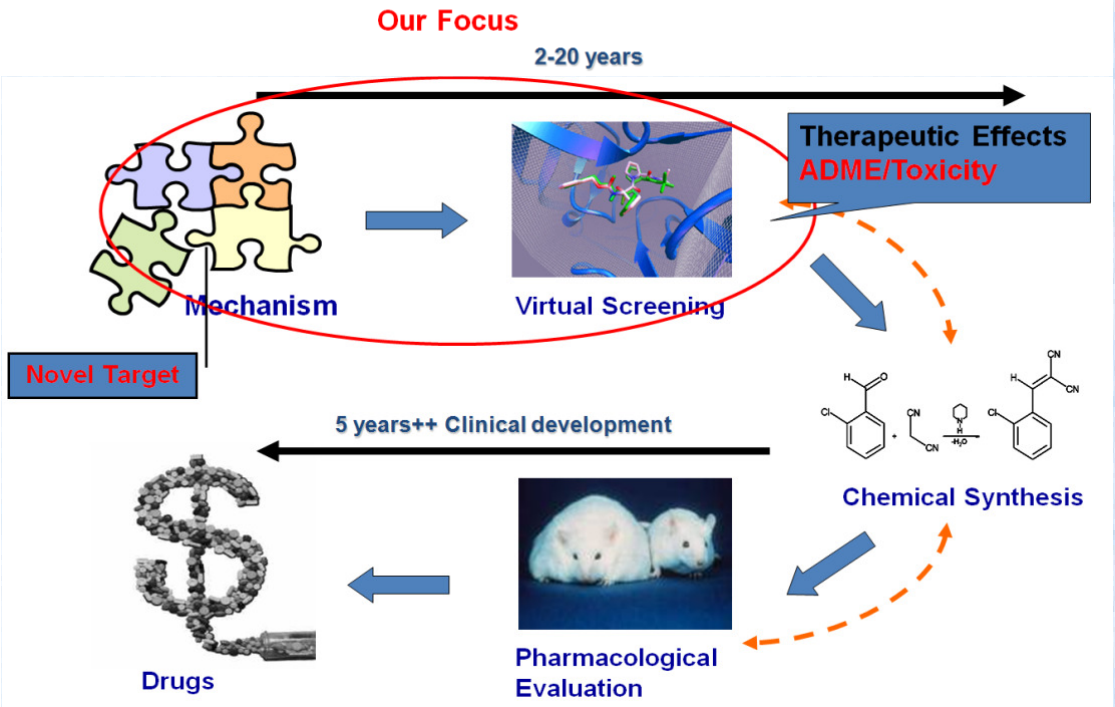
GEO - a gene expression/molecular abundance repository supporting MIAME compliant data submissions.

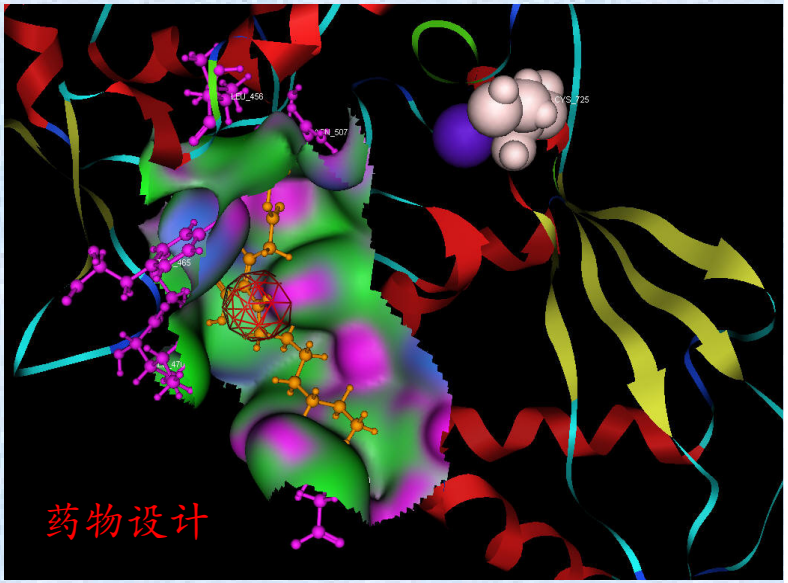


Part II: Bioinformatics-Aided Drug Design

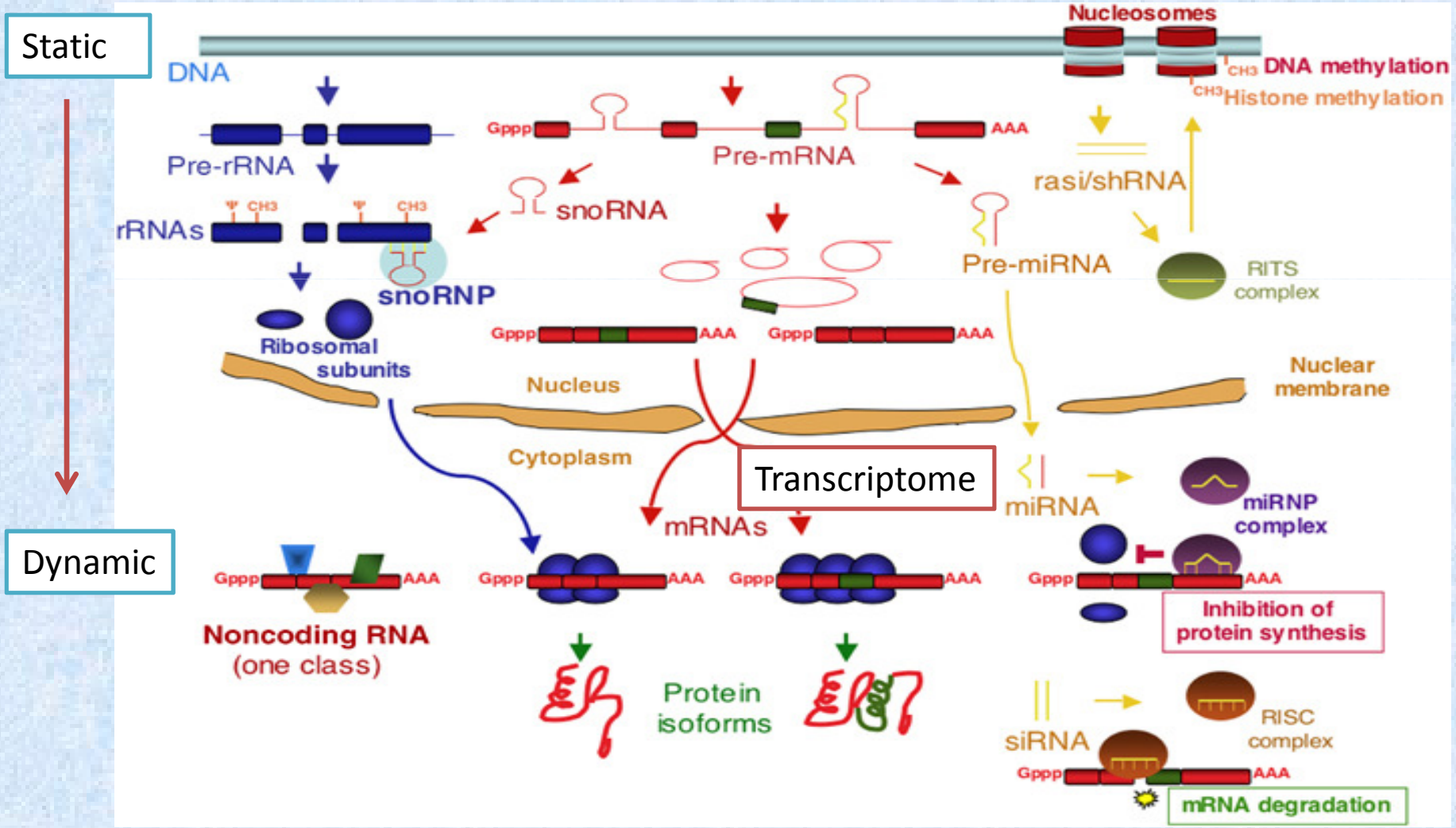
生物信息学辅助药物设计

利用数据挖掘等信息学技术开展药物药效、药代动力学和毒性等的分子机制研究；并开发整合相关软件和数据库以进行新药物靶点的发现及高通量药物筛选。



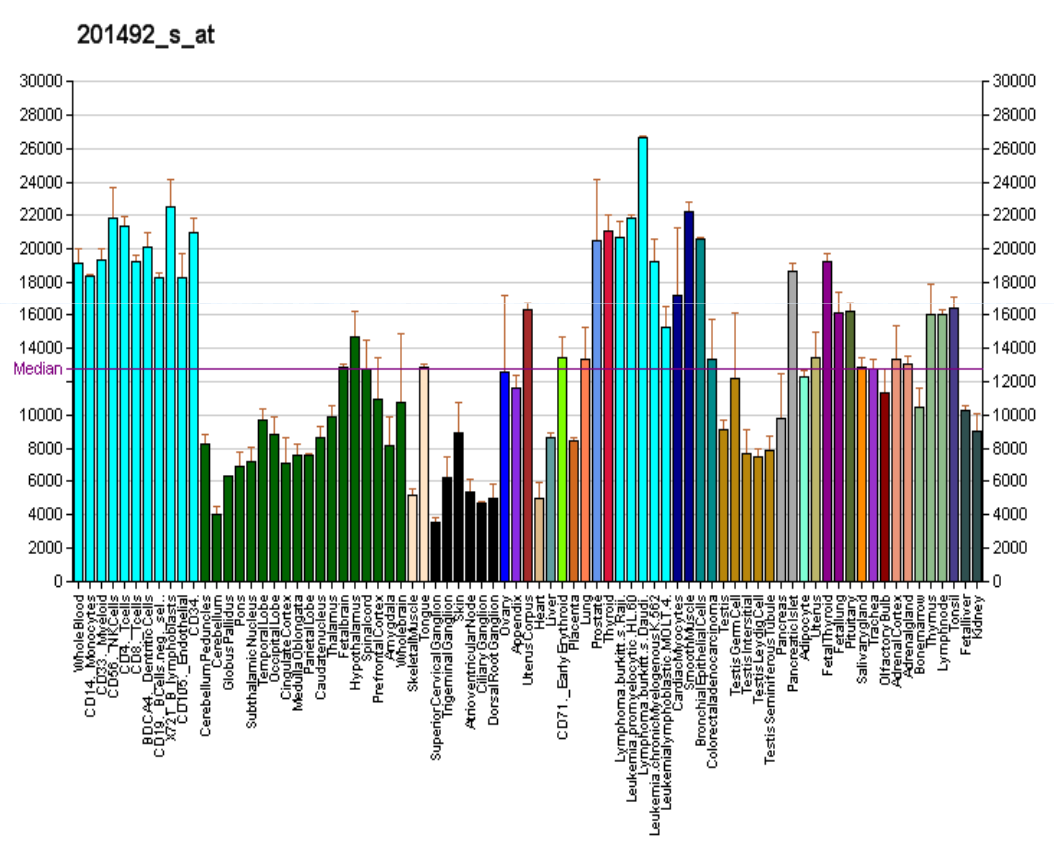


Monitoring Gene Expressions Is a Good Way to Understanding Gene Nature



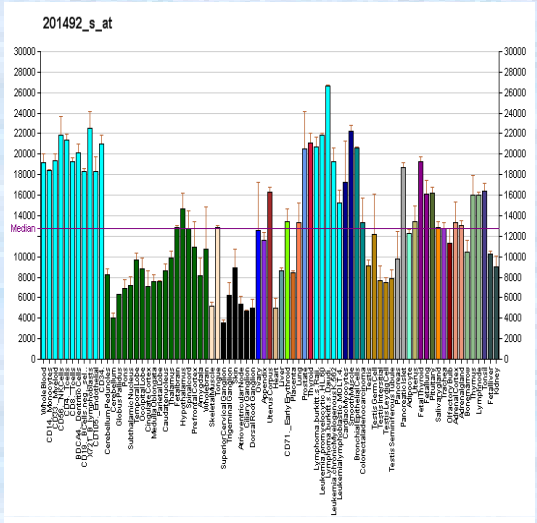
Genes Tend to Vary Their Expressions

- Under different spatiotemporal conditions
  - Tissues
  - Cell lines
  - Cell cycle
  - Developmental stages
  - Differential stages
  - …
- Under different physiological states
  - Pathogenetic periods
  - Tumorgenetic periods
  - Chemotherapy
  - …
- Under other serial conditions
  - Organisms
  - …

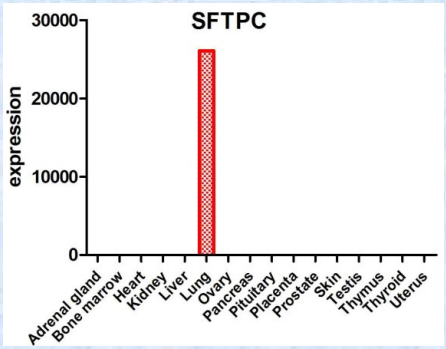


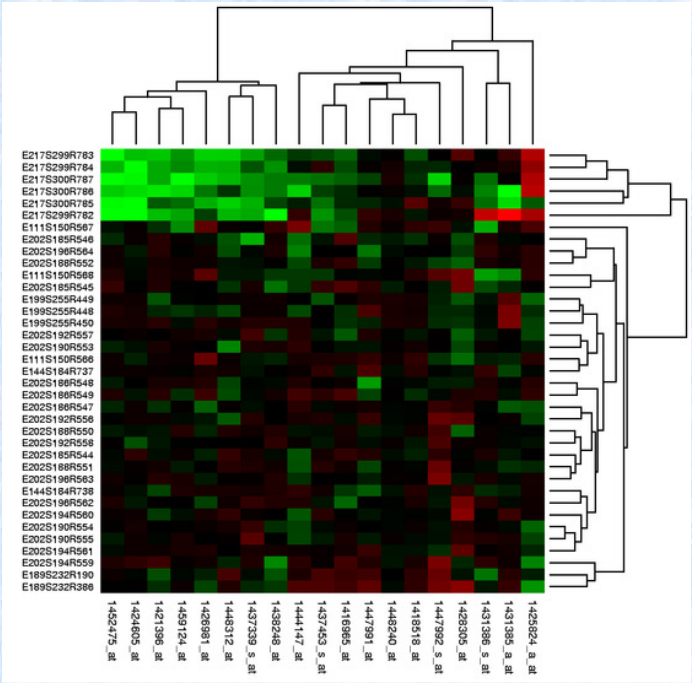
RPL41(Ribosomal protein L41 gene)

Variations of Gene Expression Carry Crucial Information



- What the gene does
- Dynamics of gene nature
- Role in physiological process
- Role in pathogenesis
- Role in different organisms
- …



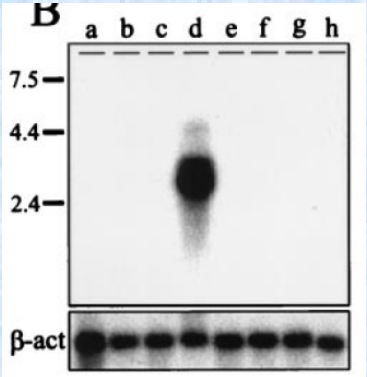


Method for Measuring Gene Expression Variations

Tranditional experiment method

- RT-PCR
- Northern blot
- Western blot
- In situ hybridization
- …

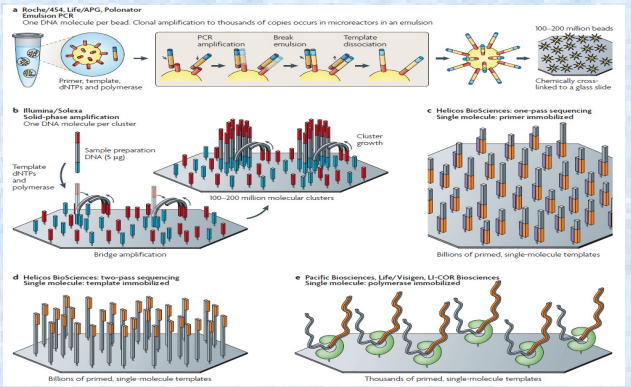
Selected genes



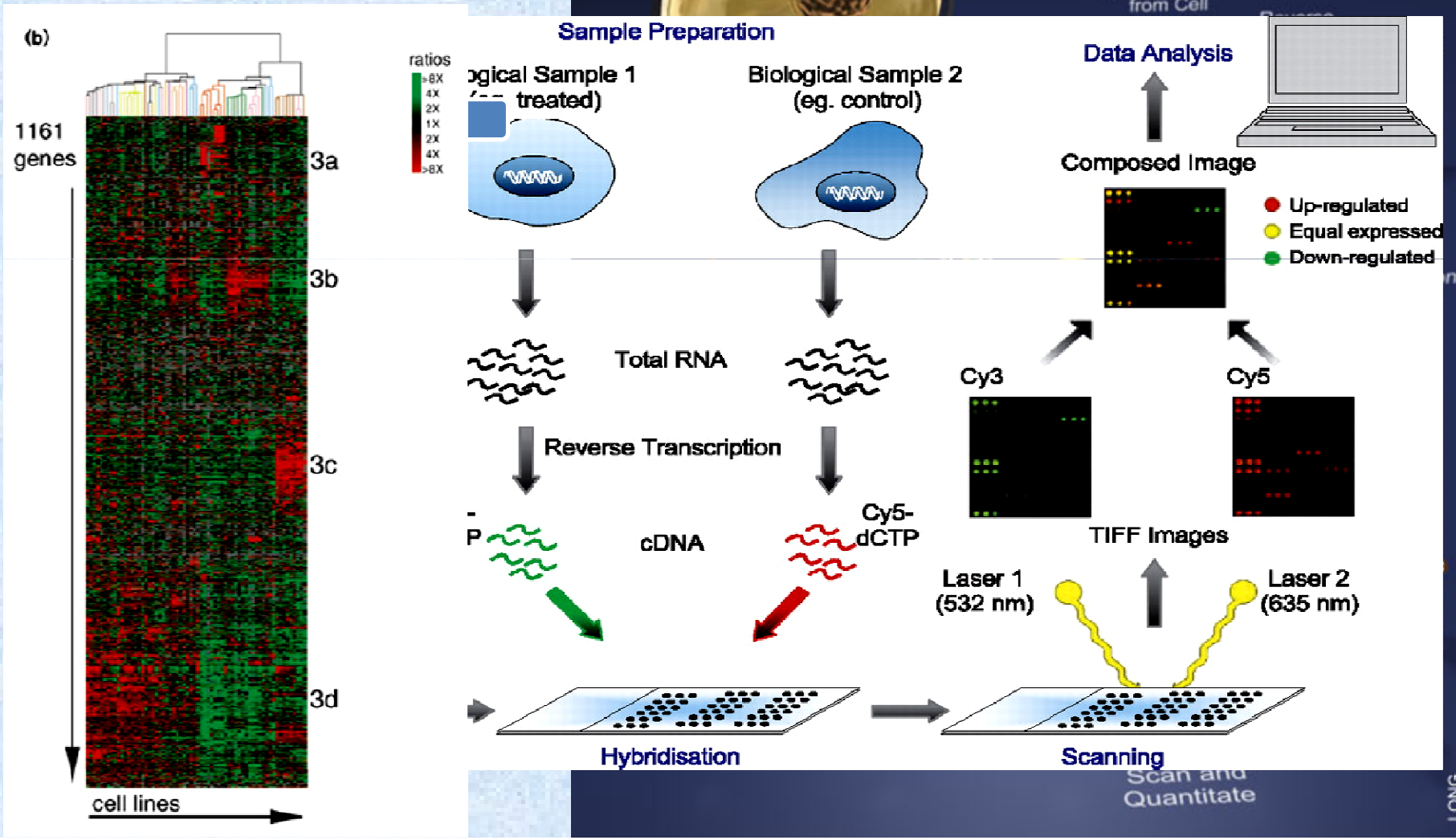
High throughput (HTP) method

- Microarray
- Serial Analysis of Gene Expression (SAGE)
- Next Generation Sequencing (NGS)
- Tilling Array

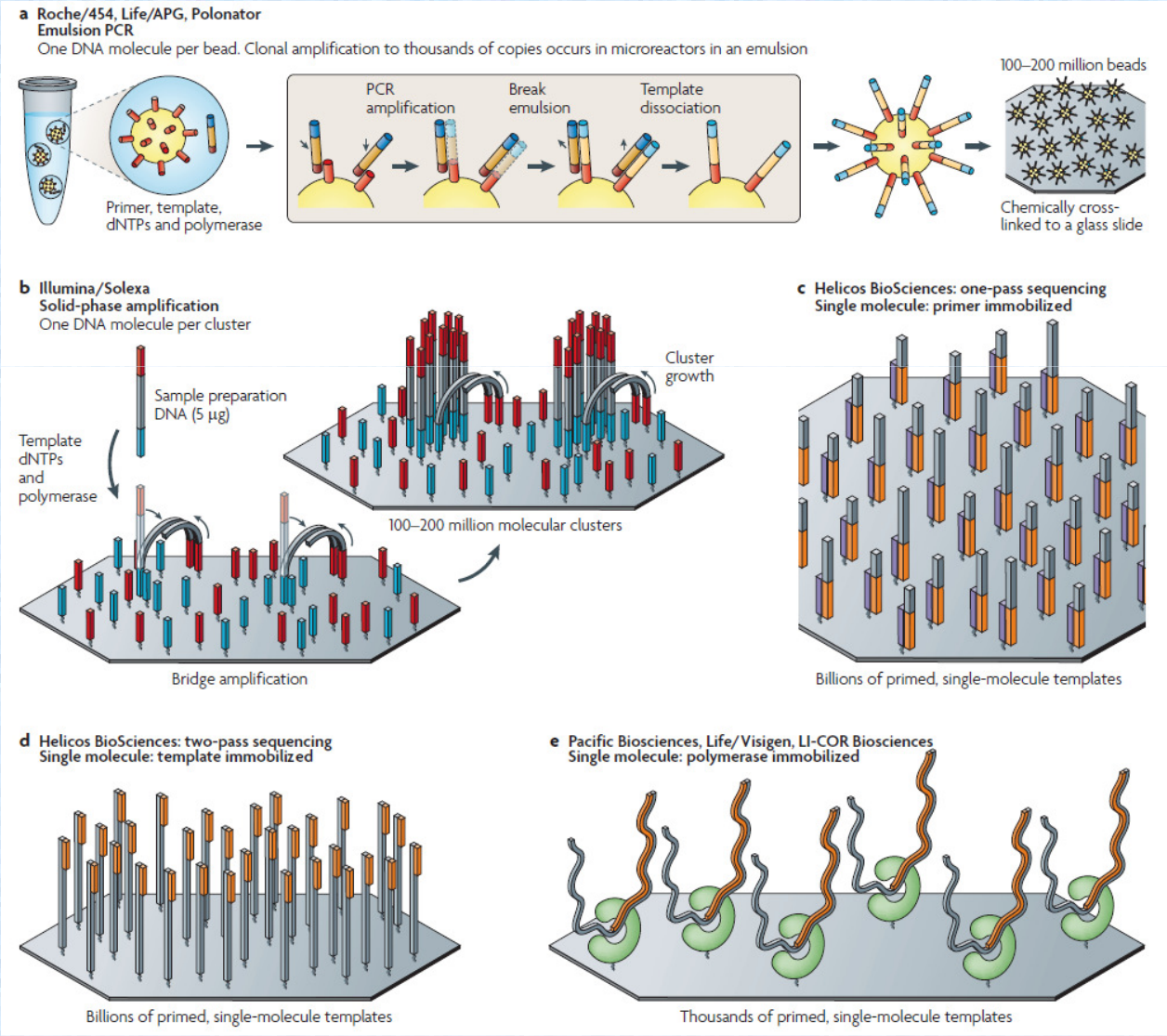
A large number of genes



Microarray

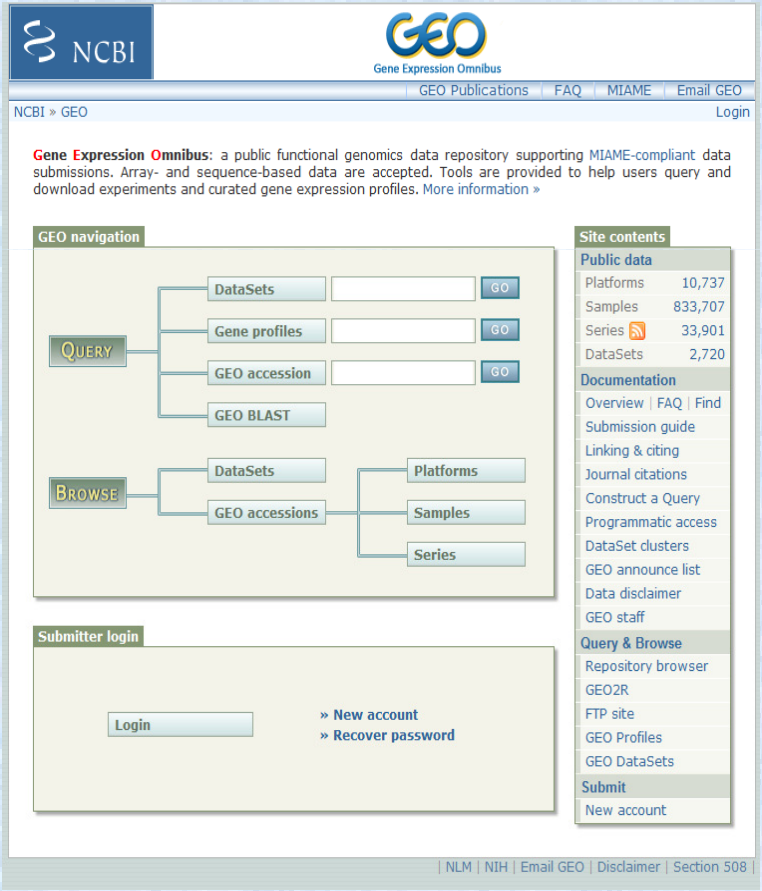

(b)
1161 genes
3a
3b
3c
3d
cell lines
ratios
Sample Preparation
Biological Sample 2 (eg. control)
Total RNA
Reverse Transcription
cDNA
Cy5-dCTP
Hybridisation
Scanning
Data Analysis
Composed Image
Up-regulated
Equal expressed
Down-regulated
Cy3
Cy5
TIFF Images
Laser 1 (532 nm)
Laser 2 (635 nm)
mRNA Extracted from Cell

Next Generation Sequencing

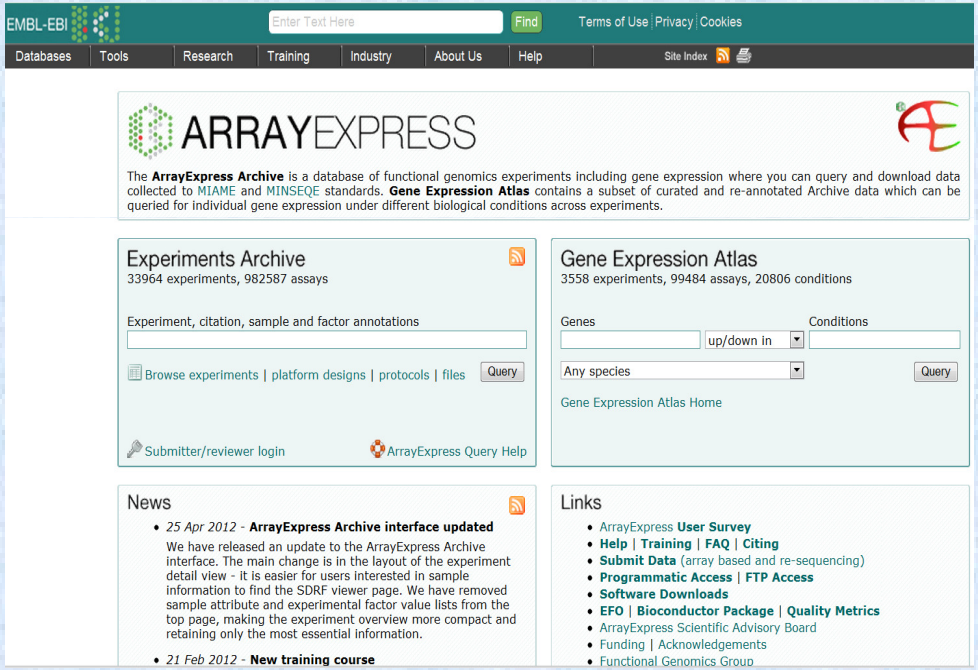


The Repositories for HTP Data

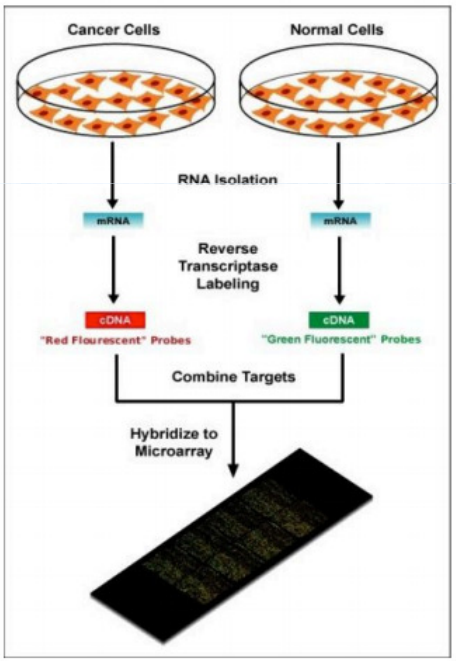
http://www.ncbi.nlm.nih.gov/geo/



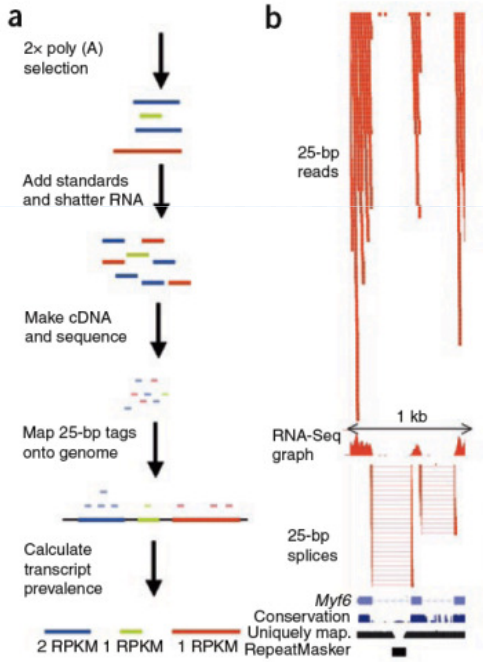
http://www.ebi.ac.uk/arrayexpress



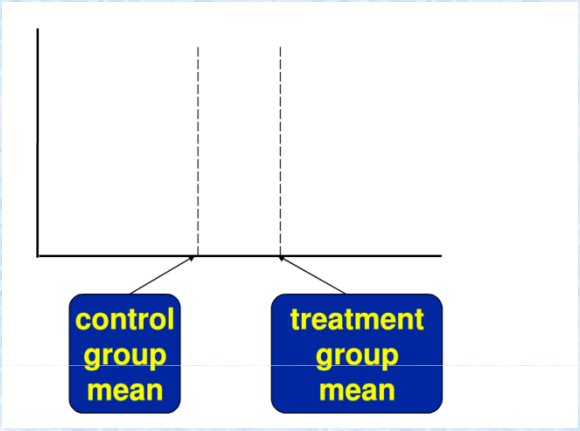
Differential Analysis



http://en.wikipedia.org/wiki/DNA_microarray



Mortazavi et al., Nature Methods, 2008



How much different between two expressions?

- **fold change, FC**
- ***t*-test**
- **analysis of variance, ANOVA**
- **nonparametric analysis**
- **regression analysis**

Comparison of Differential Analysis Tools for RNA-seq

Cuffdiff

Depends on transcript counts.

Pros.

-Easy to learn

-dependence on transcript counts make the method more reliable in terms of biology.

Cons.

-Reduced sensitivity and specificity

-Significant number of false positives in the null model

DESeq, edgeR, baySeq

Depends on raw read counts.

Pros.

-Improved sensitivity and specificity

-Good control of false positive errors

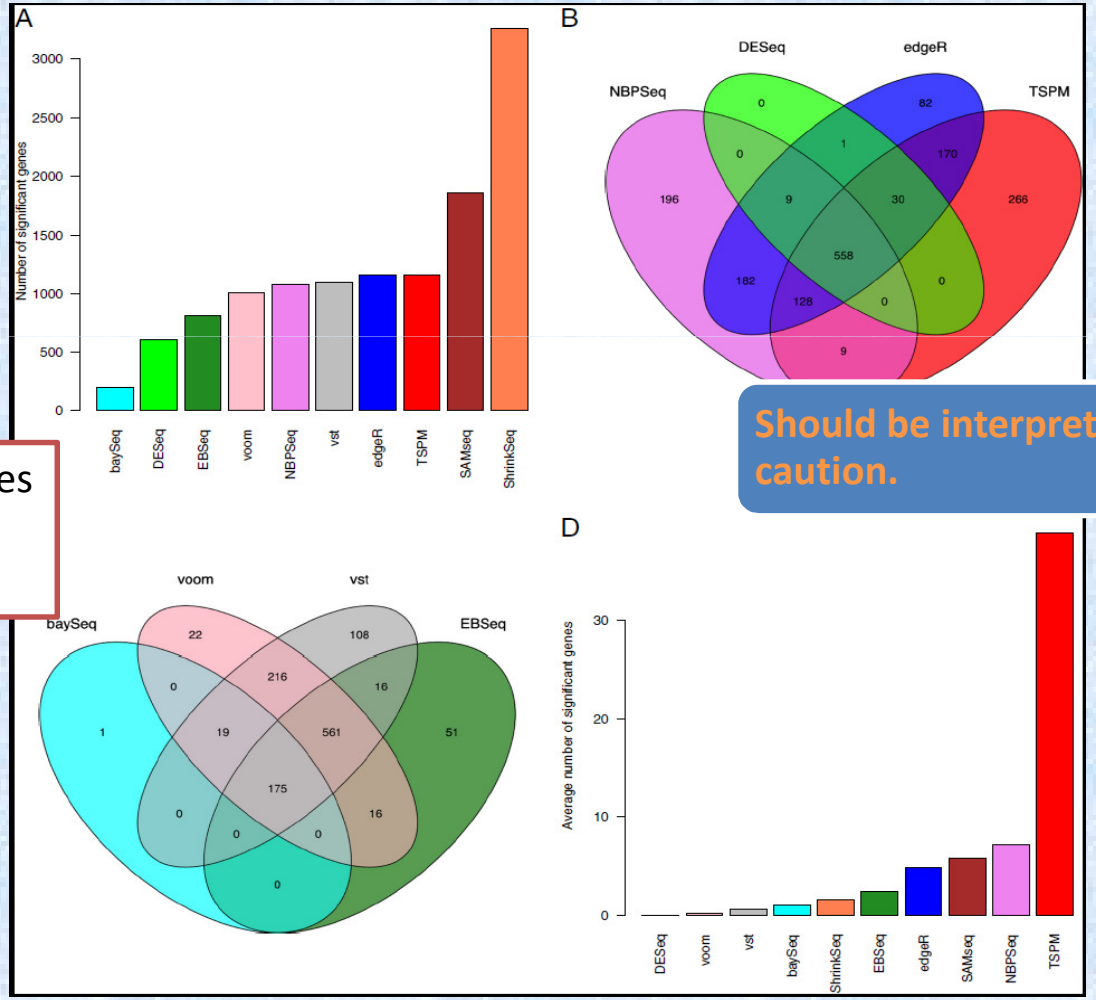
Cons.

-Kowledge of R language and Statistics is required.

-When the method is based on other distributions, negative binomial model is not a clear winner.

PoissonSeq, limma compared favorable to those.

A comparison of methods for differential expression analysis of RNA-seq data

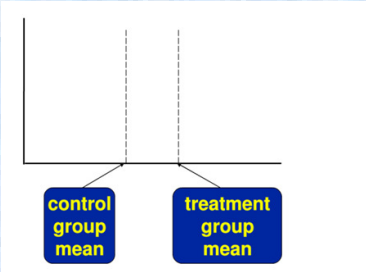


- Number of significant genes
- Significant order

Should be interpreted with caution.

Soneson and Delorenzi. *BMC Bioinformatics*. 2013 14:91

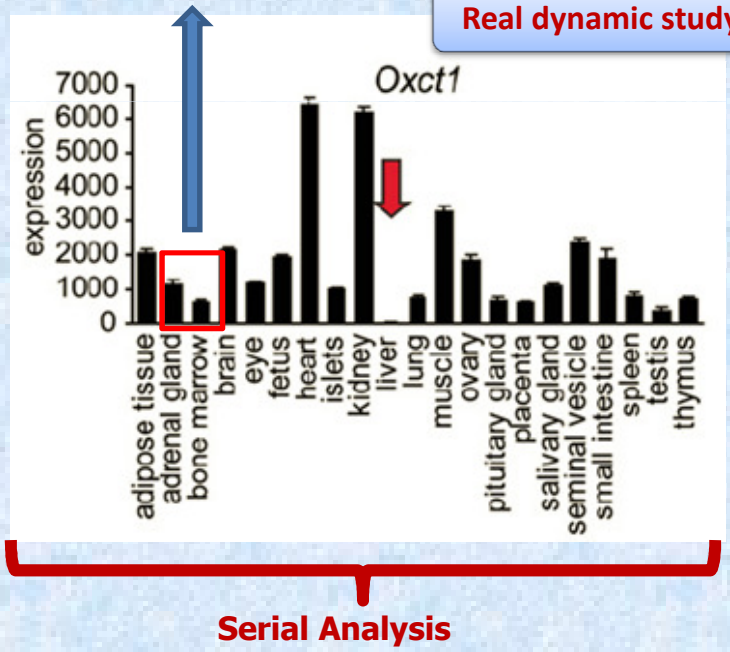
Differential Analysis Vs. Serial Analysis



-fold change, FC
-*t*-test
-analysis of variance, ANOVA
-nonparametric analysis
-regression analysis

Differential Analysis

Real dynamic study



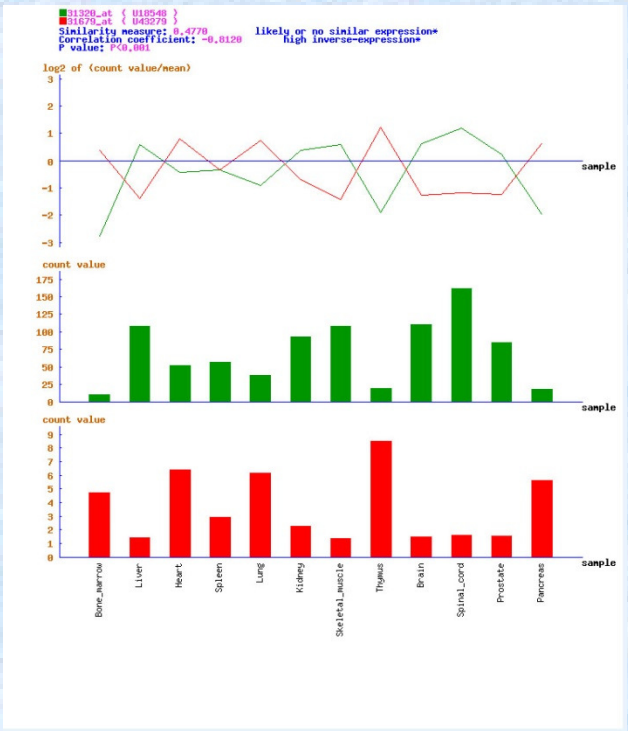
Serial Analysis

To be solved

- Is gene expression status consistent with a certain condition or is it a common phenomenon?
- How does a gene perform during the course of a condition change, such as in different developmental stages?
- Is there any connection between variably expressed genes?
- What could the regulatory mechanism underlying the different physiological conditions be?
- ...

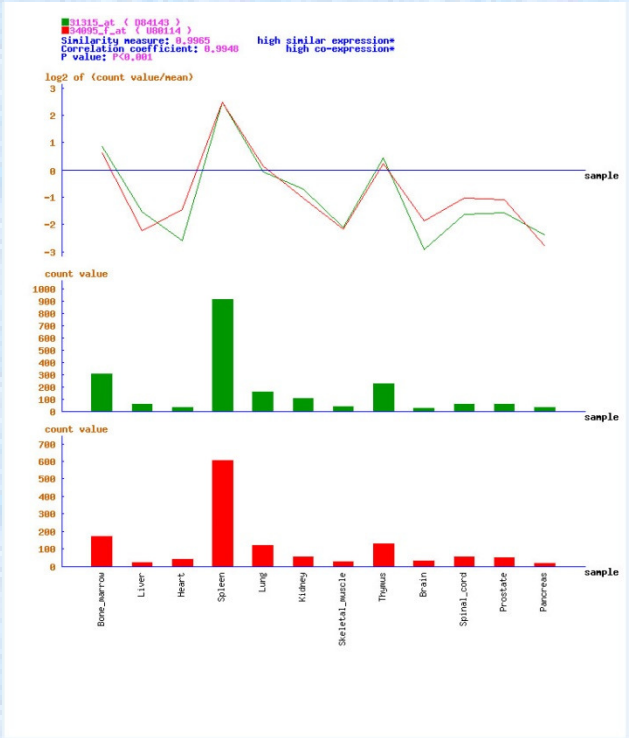
Compared to differential analysis, serial analysis could help us more globally and dynamically observe gene expression variations.

Correlation Analysis of Serial Transcriptomes



*This standard is just for suggestion!

	similar expression	co-expression	inverse-expression
high	SM>=0.95	r>=0.90	r<=-0.80
medium	0.95>SM>=0.80	0.90>r>=0.75	-0.60>=r>-0.80
likely or no	SM<0.80	0.75>r>=0	0>r>-0.60



*This standard is just for suggestion!

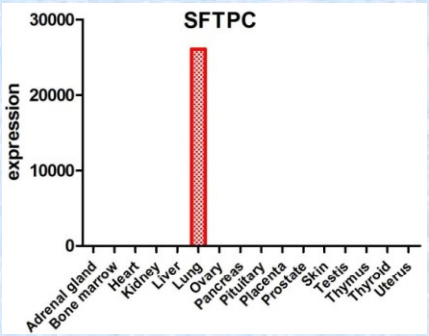
	similar expression	co-expression	inverse-expression
high	SM>=0.95	r>=0.90	r<=-0.80
medium	0.95>SM>=0.80	0.90>r>=0.75	-0.60>=r>-0.80
likely or no	SM<0.80	0.75>r>=0	0>r>-0.60

Correlation may infer the connection between two genes and their protein products: interaction or regulation.

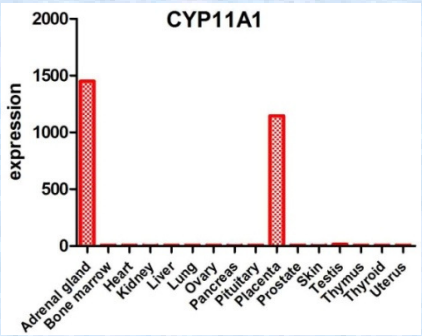
YP Wang, *et al. Nucleic Acids Research*, 2006, Vol. 34, W492-7
JB Pan, *et. al. Bioinformatics* 2012, 28(11): 1544-5

Pattern Genes Provide a Shortcut ...

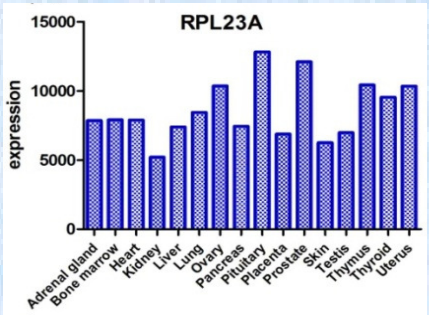
Pattern genes are genes that exhibit modularized expression behavior under serial physiological conditions.



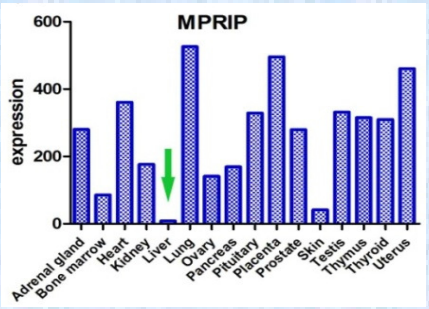
Specific Gene



Selective Gene



Housekeeping Gene



Repressed Gene

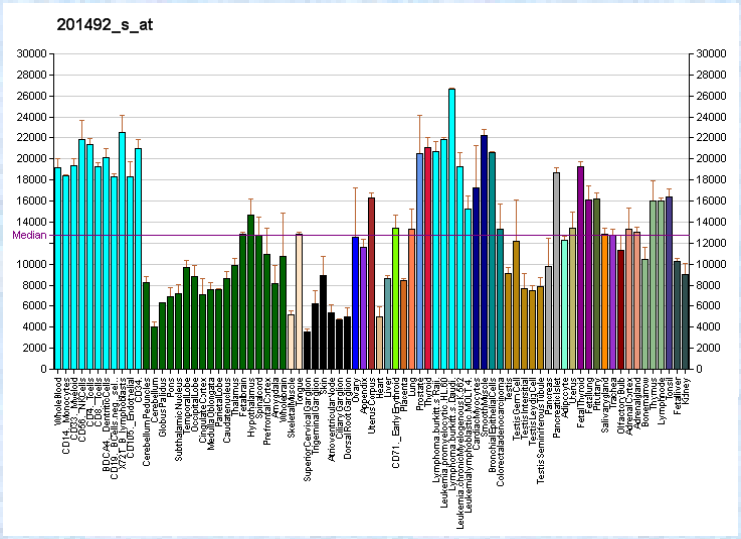
global and **dynamic** understanding of gene functions and their roles in particular biological events like development and pathogenesis.

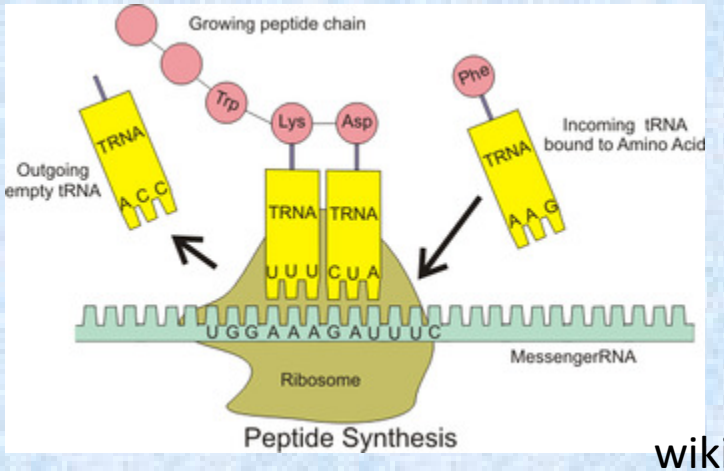
Housekeeping genes

Housekeeping genes are expressed ubiquitously and evenly under all physiological conditions.

- ✓ **Maintain basal cellular functions.**
- ✓ **Deficiencies of which will likely lead to disease.**
- ✓ **Typically adopted as molecular controls in measuring gene expressions.**

RPL41(Ribosomal protein L41)





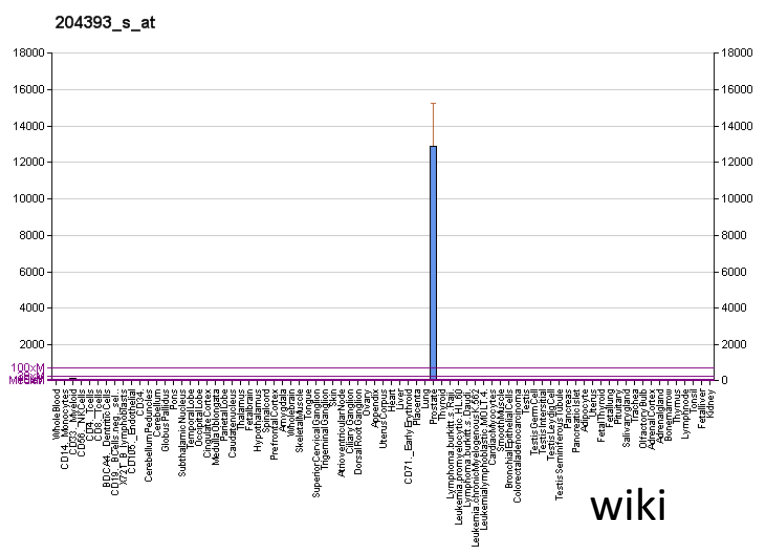
wiki

Specific/Selective genes

Specific/selective genes are those genes which are preferentially expressed under one or more conditions

- ✓ Typically considered as markers of the initiation or existence of some biological phenomena like development.
- ✓ Associated with diseases.
- ✓ Potential biomarkers and targets for disease diagnosis and treatment.

ACPP (Acid Phosphatase, prostate)



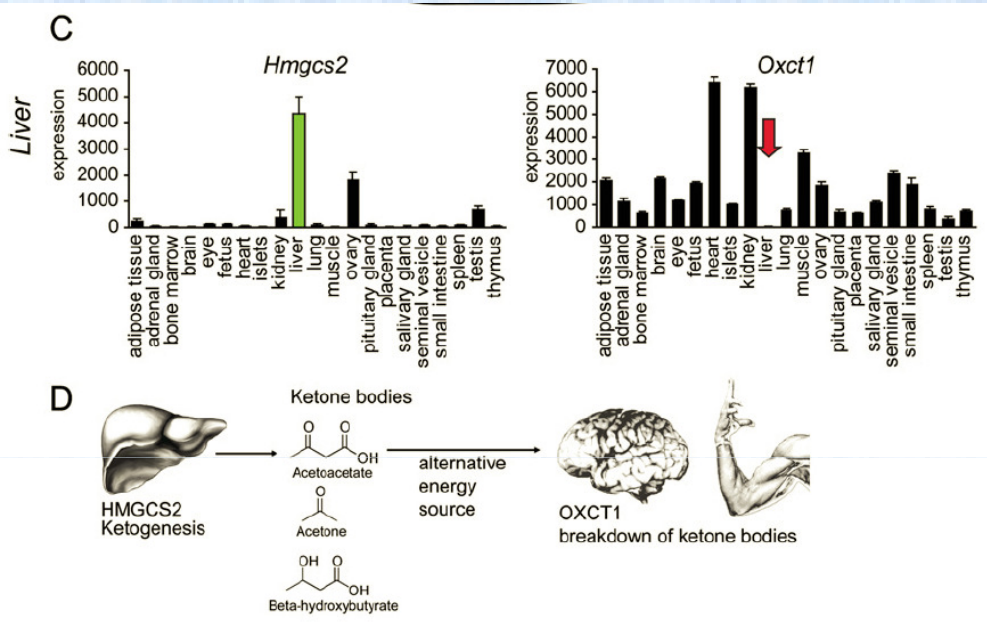
Acid phosphatase (prostate) is an enzyme produced by the prostate. It may be found in increased amounts in men who have prostate cancer or other diseases.

NCBI gene

Repressed genes

Repressed genes are expressed under almost all conditions, except for one or several conditions.

- ✓ **Repressed in the specific target condition to ensure correct function, when perturbed, can lead to diseases.**



The encoded protein by Oxct1 is an enzyme that plays a key role in ketone body degradation, which provides an alternative energy source to many tissues during fasting. Disallowance in liver seems appropriate as this tissue is specialized in ketone production in the fasted state to support metabolism of other tissues.

Genome Res 2011, **21(1): 95-105**

The Gene Expression Profiles

Probe Set	Gene Titl	Gene Symbol	Amygdala	Appendix	Cerebellu	Cerebellu	Ciliary g	Cingulate	Heart	Hypothala	Kidney	Liver
gnf1h0834	A1BG	alpha-1-B glycopro	46.25	45.7	54.3	44.6	51.6	48.4	47.85	44.05	53.75	66511.3
220951_s_	A1CF	APOBEC1 complement	858.6	1058.05	967.7	976.35	663.95	991.4	695.2	1001.6	1956.45	3857
221217_s_	A2BP1	ataxin 2-binding p	1944.1	47.85	116.7	486	35.45	1341.95	91.4	107	36.05	47.85
217757_at	A2M	alpha-2-macroglobul	4214	11173.7	1616.15	2694.05	3446.8	3035.45	5292.45	9723.65	5888.7	23423.7
gnf1h0648	A2ML1	alpha-2-macroglobul	66.1	65.05	77.45	63.95	79.05	68.8	77.95	62.9	79.05	67.75
219488_at	A4GALT	alpha 1,4-galactosy	258.6	182.25	195.2	271.5	183.85	279	307.55	293.55	187.6	323.15
218075_at	AAAS	achalasia, adrenoc	951.6	1141.95	955.4	1577.95	1100.5	1468.8	1195.7	1219.35	972.6	1178.5
218434_s_	AACS	acetoacetyl-CoA syn	1416.65	216.15	257	306.45	139.2	966.1	399.45	466.15	354.3	173.65
205969_at	AADAC	arylacetamide deac	51.6	57.55	40.85	55.4	43	60.2	50	59.1	39.25	794.1
gnf1h0617	AADACL1	arylacetamide deac	1747.85	70.95	143	72.55	78.45	796.2	74.7	661.85	88.7	72.05
202851_at	AAGAB	alpha- and gamma-a	244.05	214.5	177.45	410.2	187.1	270.95	229.55	258.6	181.7	346.25
205434_s_	AAK1	AP2 associated kin	1068.25	261.3	291.95	813.95	200.55	1691.95	142.5	309.65	280.65	196.2
201511_at	AAMP	angio-associated,	528.45	229.55	331.7	387.65	123.65	343	675.8	400.55	269.9	660.2
201000_at	AARS	alanyl-tRNA synthe	5860.2	994.6	3980.6	6839.25	1551.6	6647.3	1853.75	8987.1	2126.85	4760.75

- ✓ Irregular;
- ✓ Difficult to identify particular pattern genes;
- ✓ Can’t be well compared among genes expressed in the same sample.

Previous Methods in Identification of Pattern Genes

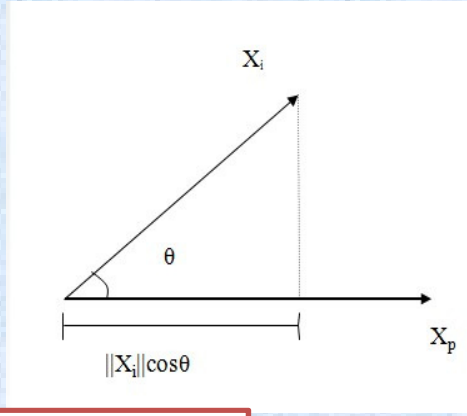
- **Cutoff**
  - Simple but qualitative;
- **relative fraction**
  - Quantitative but insensitive;
- **learning algorithms, e.g., Naive Bayes classifier and SVM**
  - Powerful but instable and hard to be implemented

Our Methodology

Probe Set	Gene Titl	Gene Symbol	Amygdala	Appendix	Cerebellu	Cerebellu	Ciliary g	Cingulate	Heart	Hypothala	Kidney	Liver
gnf1h0834	A1BG	alpha-1-B glycopro	46.25	45.7	54.3	44.6	51.6	48.4	47.85	44.05	53.75	66511.3

$$\mathbf{X_p} = (x_1, x_2, \ldots, x_i, \ldots, x_{n-1}, x_n)$$

$$\mathbf{X_i} = (0, 0, \ldots, x_i, \ldots, 0, 0)$$



Specific & Selective Gene

$$\mathrm{SPM} = \cos\theta = \frac{\mathbf{X_i} \cdot \mathbf{X_p}}{|\mathbf{X_i}| \cdot |\mathbf{X_p}|}$$

$$\mathrm{CTM}_k = \sqrt{\sum_{i=1}^{k} \mathrm{SPM}_i^2}$$

In regardless of absolute expression value.
Comparable between two profiles

Housekeeping Gene

$$\mathbf{X}_{SPM} = (SPM_1, SPM_2, \ldots, SPM_i, \ldots, SPM_{n-1}, SPM_n)$$

$$\mathrm{DPM} = \sqrt{\frac{\sum_{i=1}^{n} (\mathrm{SPM}_i - \overline{\mathrm{SPM}})^2}{n-1}} \cdot \sqrt{n}$$

A transform of standard deviation and normalization.
Comparable between two profiles.

Repressed Gene

$$\mathrm{RPM}_k = \frac{\mathrm{SPMrep\_max}}{\mathrm{SPMexp\_min}}$$

SJ Xiao, *et. al. Bioinformatics* 2010, 26(9): 1273-5
JB Pan, *et. al. Bioinformatics* 2012, 28(11): 1544-5

Quantitative Identification of Pattern Genes

$$SPM = \cos\theta = \frac{\mathbf{X}_i \cdot \mathbf{X}_p}{|\mathbf{X}_i| \cdot |\mathbf{X}_p|}$$

Specific Gene

SPM > 0.9

$$CTM_k = \sqrt{\sum_{i=1}^{k} SPM_i^2}$$

Selective Gene

$2 \le k \le 6$, $SPM_{i\ (1\ to\ k)} > 0.3$, and $CTM_k > 0.9$

$$DPM = \sqrt{\frac{\sum_{i=1}^{n}(SPM_i - \overline{SPM})^2}{n-1}} \cdot \sqrt{n}$$

Housekeeping Gene

Ubiquitously express
DPM < 0.3

$$RPM_k = \frac{SPMrep\_max}{SPMexp\_min}$$

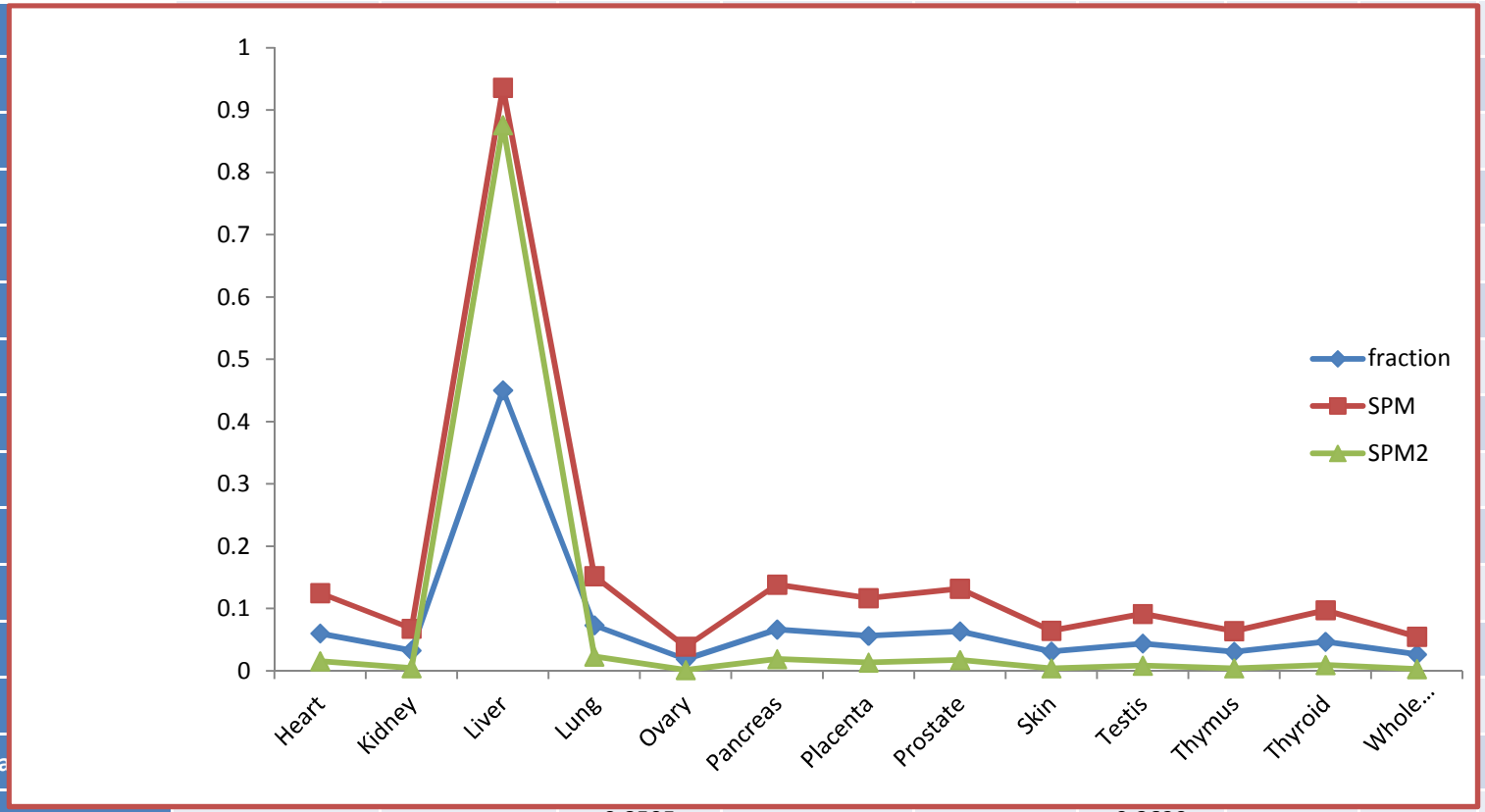
Repressed Gene

$1 \le k \le 6$, $SPM_{i\ (1\ to\ k)} < 0.1$, and $RPM_k < 0.2$

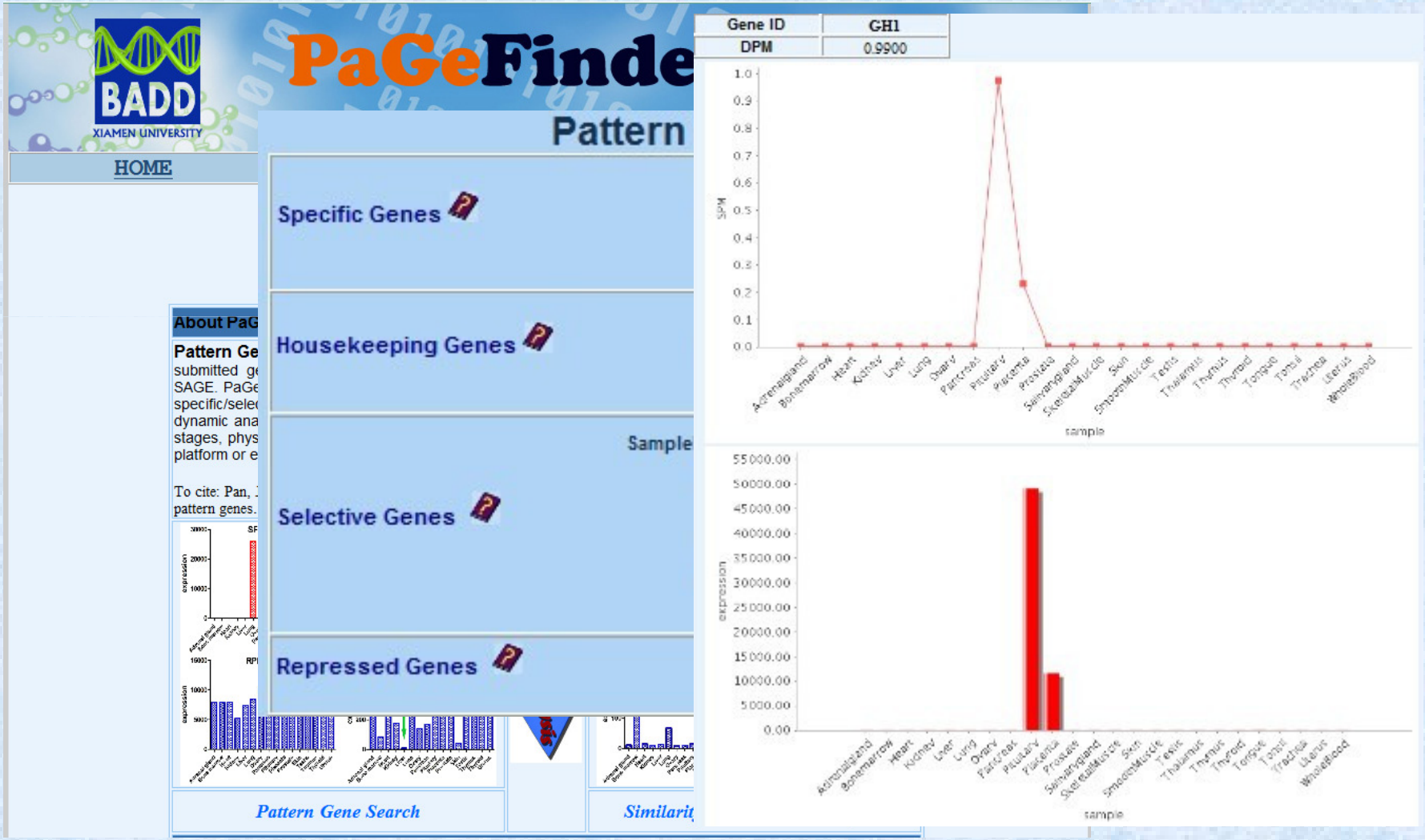
Jian-Bo Pan, *et. al. Bioinformatics* 2012, 28(11): 1544-1545

Comparison to Other Methods

Gene/ProbesetID / Tissue	STEAP3 / 218424_s_at			ETFB / 202942_at			RPL23A / 208825_x_at		
	value	Fraction	SPM	value	Fraction	SPM	value	Fraction	SPM
Heart									0.3523
Kidney									0.1722
Liver									0.3111
Lung									0.2836
Ovary									0.3492
Pancreas									0.2378
Placenta									0.2082
Prostate									0.3806
Skin									0.2075
Testis									0.2141
Thymus									0.2774
Thyroid									0.2904
Whole brain									0.2284
Standard Devia									0.0651
DPM			0.8505			0.8620			0.2346



http://bioinf.xmu.edu.cn/PaGeFinder/index.jsp



厦门大学生物信息学辅助药物开发研究组

Bioinformatics-Aided Drug Discovery Group, Xiamen University

http://bioin

Gene information		
Gene Symbol	Gene Description	Entrez Gene ID
CYP3A4	cytochrome P450, family 3, subfamily A, polypeptide 4	1576

Evidences		
Dataset	ID	GDS3113|Array_count
	Title	Various normal tissues
	Summary	Analysis of various normal tissues. Results provide insight into housekeeping and tissue-specific genes.
	PUBMED	Dezso Z, Nikolsky Y, Sviridov E, Shi W et al. A comprehensive functional analysis of tissue specificity of human gene expression. BMC Biol 2008 Nov 12;6:49. PMID:19014478
Profile	Probe Set Id:	103888
	Total Sample Numbers	26
	Profile Variation	DPM=0.9855
Specific Gene	Sample|SPM	liver|SPM=0.9449
Selective Gene	CTM2	0.9998
	Sample|SPM	liver|SPM=0.9449 small intestine|SPM=0.3267

thaliana

itis elegans

nelanogaster

us

egicus

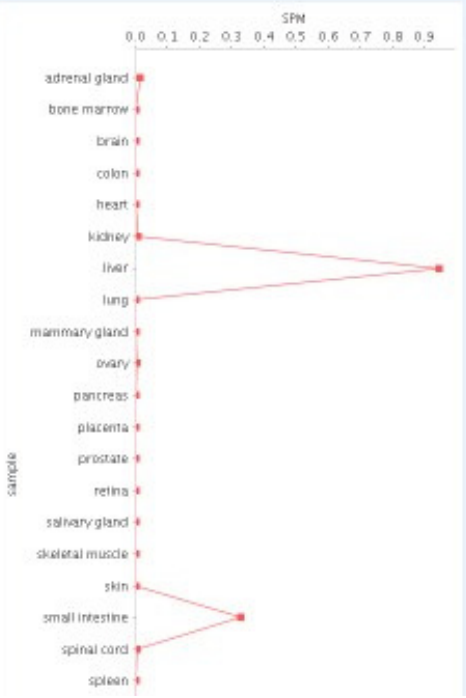
ces cerevisiae

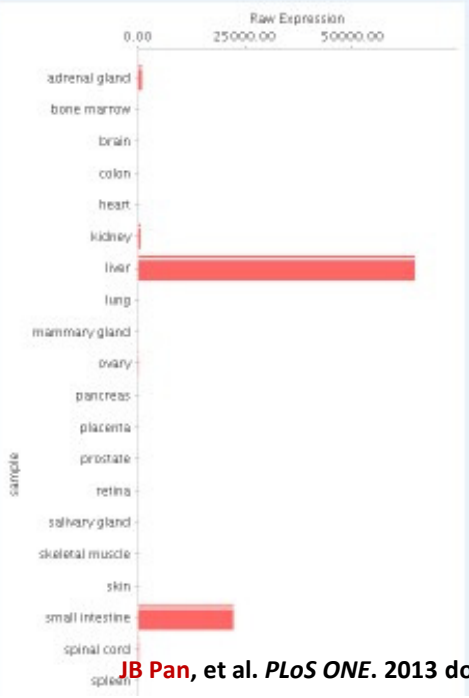
HOME SEARCH

Gene Symbol	
CYP3A4	GDS31
CYP3A4	BIOGPS_U
CYP3A4	GDS10
CYP3A4	GDS181_T
CYP3A4	

Tips

Classes		
	Selective Genes	Repressed Genes
	Y	–
	Y	–
	–	–
	–	–
	–	–





11 model organisms

143 quality microarray

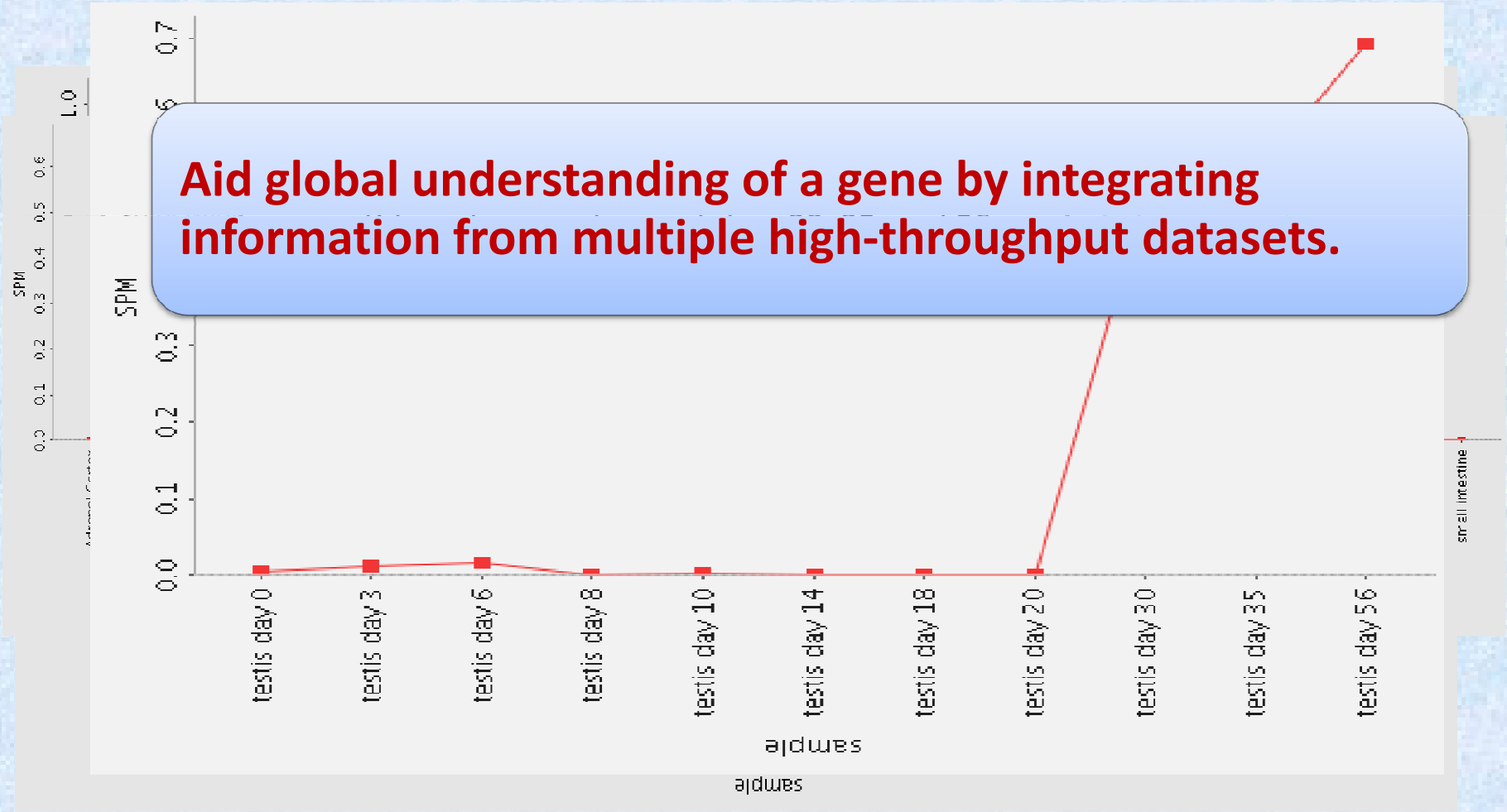
1,145,277 profiles

JB Pan, et al. *PLoS ONE*. 2013 doi:10.1371/journal.pone.0080747.

Potential Applications

- Dynamic and global understanding of gene expression
- Biomarkers for prognosis.
  - Specific/selective genes
  - Repressed gene
- Therapeutic targets for new drug discovery
  - Specific/selective genes
  - Repressed genes
- Molecular controls
  - Housekeeping genes (positive control)
  - Repressed gene (negative control)
- Further biomedical research

Protamine 1 gene (PRM1) encodes a small, arginine-rich, nuclear protein that replaces histones in the chromatin of sperm during the haploid phase of spermatogenesis.

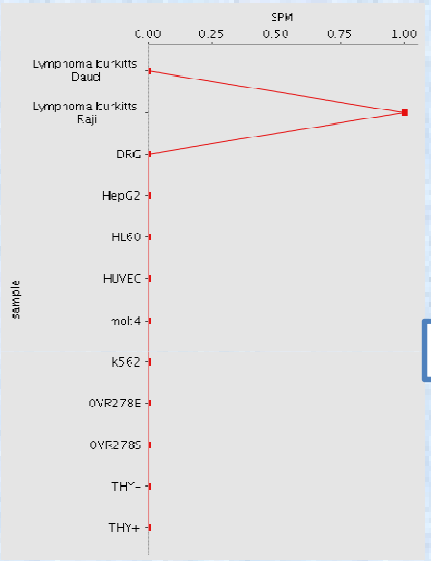


Aid global understanding of a gene by integrating information from multiple high-throughput datasets.

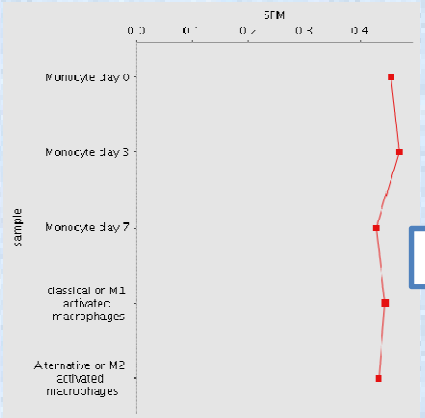
Selection of Reference Gene

GAPDH

- ✓ A reference gene should be ubiquitously expressed under all studied conditions.
- ✓ It should have relatively high expression levels so that it can be easily and stably detected.
- ✓ Its expression should be insensitive to condition changes.
  - ✓ GAPDH is selectively expressed in skeletal muscle ($SPM_{skeletal\ muscle} = 0.47$).
  - ✓ GAPDH expression is nearly unchanged in some cell types.



Cells



Development

Conclusions

- Quantitative identification of pattern genes would be good point of penetration in understanding serial transcriptomes.
- Pattern genes provide a shortcut to study dynamic nature of gene.
- Selective genes mediate cross-talks between tissues.
- Specific/selective genes are often associated with tissue-specific pathogenesis.
  - They are potential biomarkers.
  - They are potential therapeutic targets.

Acknowledgements

- Asst Prof. Quan Zou 邹权
- Jian-Bo Pan 潘建波
- Shi-Chang Hu 胡始昌
- Mei-Chun Cai 蔡梅春
- Yin-Bo Li 李银波
- Hai-Jing Ji 金海晶
- Ke Liu 刘珂
- Quan Xu 徐全
- Dan Shi 石丹
- Yan-Mei Qin 覃杨梅